

Award Number: W81XWH-06-1-0761

TITLE: Surgical Technology Integration with Tools for Cognitive Human Factors (STITCH)

PRINCIPAL INVESTIGATOR: W. Brent Seales, Ph.D.

CONTRACTING ORGANIZATION: University of Kentucky Research Foundation  
Lexington, KY 40506-0057

REPORT DATE: October 2008

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command  
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

☒ Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. <b>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</b>					
1. REPORT DATE (DD-MM-YYYY) 24-10-2008		2. REPORT TYPE Annual		3. DATES COVERED (From - To) 25 Sep 2007 - 24 Sep 2008	
4. TITLE AND SUBTITLE  Surgical Technology Integration with Tools for Cognitive Human Factors (STITCH)				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER W81XWH-06-1-0761	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) W. Brent Seales, Ph.D.  Email: seales@uky.edu				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)  University of Kentucky Research Foundation 109 Kinhead Hall Lexington, Kentucky 40506-0057				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)  U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S) USAMRMC	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The need for accurate assessment in surgical training has become even more important with the development of new surgical technologies, many of which have transformed methods of treatment for both the patient and the surgeon. Difficult-to-master technologies such as the components of Minimally Invasive Surgery (MIS) highlight the need for surgical competence but do not inherently provide a solution for how to define and measure it. The long-term goal of this research is to build an integrated surgical technology environment designed for the continuous monitoring of task performance, with a particular focus on the inclusion of important but currently-overlooked cognitive measures.					
15. SUBJECT TERMS  None provided.					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT  Unlimited	18. NUMBER OF PAGES  14	19a. NAME OF RESPONSIBLE PERSON W. Brent Seales
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (include area code) (859) 257-3063

Introduction .....	4
Research Accomplishments.....	4
1. Cognitive Ergonomics.....	4
2. Clinical Assessment at the UMMC MASTRI Center.....	8
3. Tools and Display Technology.....	8
4. Supported Personnel.....	13
Key Research Accomplishments (Summary) .....	13
Reportable Outcomes .....	14
Conclusion.....	14

# Introduction

The need for accurate assessment in surgical training has become even more important with the development of new surgical technologies, many of which have transformed methods of treatment for both the patient and the surgeon. Difficult-to-master technologies such as the components of Minimally Invasive Surgery (MIS) highlight the need for surgical competence but do not inherently provide a solution for how to define and measure it. The long-term goal of this research is to build an integrated surgical technology environment designed for the continuous monitoring of task performance, with a particular focus on the inclusion of important but currently-overlooked *cognitive* measures.

Evaluation of surgical skill in MIS can be made more accurate, objective, and general by considering cognitive and environmental factors such as mental workload, stress, situation awareness, and level of comfort with complex tools. To date, our research has shown that a comprehensive framework for measuring cognitive human factors in MIS settings provides an important, statistically significant set of (largely overlooked, in this domain) non-redundant metrics for evaluating performance in the context of new technologies, tasks, and learning methodologies.

Software development efforts are producing a general-purpose Plug-and-Play (PnP) framework and application-specific tools usable in that framework. Well-defined methodology is being incorporated in the development process to insure required safety, reliability and robustness attributes for the problem domain. The human studies used to adapt, select, and validate the cognitive measures are divided into four parts: an *equivalence testing* phase followed by three *validity studies*. The equivalence study ensures reliability of test results after minor modifications. The subsequent validity tests assess *construct validity* (the extent to which our measures of stress and workload dissociate from measures of performance in situations where, theoretically, they should); *concurrent validity* (sensitivity to differences among surgical conditions known to differ substantially in difficulty); and *predictive power* (degree to which measures can predict more complicated indices of surgical proficiency, including adaptive aspects of performance on which students have not received explicit training).

The STITCH project has made significant progress in the development of specifications, designs, and implementations of an integrated surgical training and assessment framework and is providing assessment results for specific cognitive measures, including validity and predictive studies. These results are useful for implementing improvements in training methods that seek to use valid cognitive measures as part of the assessment strategy.

## Research Accomplishments

### 1. *Cognitive Ergonomics*

During the first quarter of year two we received the FaceLab eye tracking hardware and software. Representatives from the company traveled to Lexington and completed a full day of training on the system. The system has been integrated into both our scalable display framework and data capture evaluative environment.

**Installation and Setup:** The FaceLab eye tracking system was delivered, installed and configured on site at the University of Kentucky. This system includes an integrated software development kit (SDK) that our developers are using for custom control of the tracking environment. Representatives from FaceLab provided one full day of training to our technical team and technical leads. This training led to a complete configuration of the system and provided our team with the technical expertise necessary to integrate the device into our experimental framework.

**Trial Subjects:** After training we established baseline experience with the FaceLab system using a set of trial subjects. These subjects provided basic testing with eye calibration, screen-to-object matching and basic validation testing of calibration quality, quickness, and convenience. This also provided training for the technical team in configuring and using the system.

**FaceLab Evaluation:** The evaluative data from the FaceLab system training and trials was collected to inform the team as to how best to use the device as part of the assessment framework for current and planned studies.

During the last quarter of 2007 and the first quarter of 2008, we continued to assess the different versions of our battery of subjective mental workload measures (NASA-TLX, MRQ, and SSSQ). We focused on comparing the traditional printed format of NASA-TLX to the auditory-vocal ("hands-free") format. Our findings on the equal

sensitivity of the vocal format to task and training effects as compared to the traditional paper and pencil format has led us to move forward to develop and refine the VisRecognizer, a voice recognition based system. In the past three months, we have also begun establishing which of several time estimation metrics is most sensitive for our purposes, and we have moved forward with the automation of both the presentation and analysis functions of the time estimation task.

*Reliability assessments: Equivalent- forms reliability of printed and spoken versions of the NASA-TLX*

Sixty-four students (38 males, 26 females) were recruited from introductory Psychology classes based on responses on a screening questionnaire indicating that they were 1) frequent players of video games and were 2) in a pre-med or other healthcare curriculum. Half of the participants were randomly assigned to the auditory-vocal condition and half of them were assigned to the printed condition.

We asked participants to perform three training tasks that are among the initial tasks used at the minimally –invasive surgery training center at the University of Kentucky. These tasks were: the cannulation task, cobra rope passing task, and the peg transferred task. After orientation, participants practice each task once in open view and began a series of 90-second trials. The NASA-TLX was administered, along with a battery of other subjective measures, after each trial.

- *Correlational Analysis*

Initially, a correlational analysis was performed between vocal and written versions of the NASA-TLX, as well as between weighted and unweighted calculations of global workload. These correlations are calculated based on the mean values summarized across research participants to yield estimates for each of the 5 (block) X 3 (task) factors for each of the administration conditions. The correlation between the two administration formats for the unweighted and weighted versions was  $r = .81$  and  $r = .85$ , respectively. We consider these correlations to be acceptable, especially in light of the two formats being administered to different sets of research participants (i.e., this was not a within-subjects design due to the time constraints we had when dealing with each participant).

Correlations between the unweighted and weighted scale scores (both scores calculated for each participant) yielded a correlation of  $r = .91$ . This is consistent with previous research comparing the two methods of calculating global workload from the component scales. Moroney et al. (1992) and Byers, Bittner, and Hill (1989) obtained coefficients of equivalence of  $r = .94$  and  $r = .98$ , respectively.

- *Comparison of Sensitivity*

Although we found reasonably high correlations between the vocal and written versions of the NASA-TLX, we felt that it was also important to evaluate the relative sensitivity of the two administration techniques to variation in workload that would be predicted on the basis of task differences (cannulation vs. ring transfer vs. cobra rope) and practice (over five blocks of trials). There is likewise a concern for the weighted versus unweighted composite scores as others (e.g., Liu and Wickens, 1994) have obtained slightly larger effect sizes with the weighted version.

Independent 2-way (task X block) ANOVAs were calculated for both weighted and unweighted NASA-TLX scores for participants in the vocal and written conditions. Vocal administration resulted in between-task effect sizes (omega-squared) of .13 for both unweighted and weighted composites. For practice, effect sizes were .05 and .03 for unweighted and weighted calculations.

Turning to the written administration condition, between-task effect sizes were .098 and .13 for the unweighted and weighted versions, respectively. For variation across trial blocks, the effect sizes were .03 for both composites. These data indicate that there appear to be minimal differences in sensitivity between the 2 (weighting) X 2 (administration format) conditions (i.e., results of a mixed-factor ANOVA yielded no reliable differences).

- *Implications*

On the basis of these findings, we are pursuing the inclusion of a vocal version of the NASA-TLX (including voice-activated responding) in our current battery of cognitive ergonomics metrics for evaluating user responses to laparoscopic technology innovations. In addition, because of the limited time most surgical personnel are available for testing, we also recommend that the paired-comparison procedure for developing participant-specific scale weightings be reconsidered. In our view, there is limited payback for the extra time required to obtain the necessary information to create these weighted composites. A simple, unweighted mean of the six scale scores may be sufficient for most purposes.

### *Reliability assessments: Voice recognition software program*

During the past year, we focused on implementing and evaluating an automated, “hands-free” auditory version of several subjective mental state questionnaires. These questionnaires include NASA Task Load Index (NASA-TLX), the Multiple Resources Questionnaire (MRQ, another workload instrument that pinpoints the nature of the mental demand), and the Short Stress State Questionnaire (SSSQ, which measures self-perceptions of worry, engagement, and distress).

Our goal is to continue the development of a vocal format as an alternative to the traditional paper and pencil administration methods which are not particularly useful for evaluating surgical technologies during either simulated or real surgeries because the surgeon’s hands are already engaged. Earlier data obtained from testing 64 non-surgeons revealed that translation of the technique from visual-manual to auditory-vocal modalities would not, in principle, introduce additional error variance or systematic biases. We have described in more detail this reliability assessment in the above section and we have submitted these data to the Human Factors and Ergonomics Society’s 2008 conference. The paper has been accepted for inclusion.

The conducted intensive studies of the reliability of our modified administration techniques for our cognitive measures compared to traditional methods. Our modifications are necessitated by the relatively long, continuous, and manually-intensive nature of surgical tasks, making traditional paper-and-pencil formats inappropriate. Note that because of the relatively large number of subjects needed for initial psychometric evaluations, we have been using participants from the general undergraduate population, and they have been using trainer boxes situated in labs at the UK Dept. of Psychology and UK’s Center for Visualization and Virtual Environments. During the year we began administering our secondary task and subjective state measures (i.e., NASA-TLX, MRQ, SSSQ) to medical students at UK’s Center for Minimally-Invasive Surgery. Specifically, we have 1) replicated our equivalent-forms reliability study of the vocal and standard paper-and-pencil subjective state measures on medical students in the actual training environment, and 2) we have used the measures in an evaluation of the impact of 3D enhancements afforded by the DaVinci surgical robot on initial attempts by medical students to use the robot.

### *Assessing the Performance of Medical Students*

We spent much of late May and early June creating a research station inside UK’s MIS Center (a training facility at the UK Candler Medical Center) in which we can test medical students, residents, and surgeons. Now completed, the station will be available for the duration of the STITCH project to allow testing of research participants in a more ecologically valid environment. The new station includes a calibrated 6-projector tiled display system that allows the presentation of large-scale displays (up to 120” diagonal), as well as the presentation of multiple simultaneous information sources (e.g., simulated pre-op imagery, a performance dashboard). The station also includes a standard Stryker box trainer, assorted surgical instruments, a variety of training materials to be manipulated by participants (e.g., small rings, ropes, tubes), and a computer for automated stimulus presentation and data collection.

We have completed data collection in our first study at the MIS Center, which involved a replication of our Fall 2007 equivalent-forms reliability study. In both studies, we used the comparison of subjective states across three training tasks (cannulation, ring transfer, and cobra rope) as a means of evaluating the sensitivity and comparability of mental workload and stress measurements taken using different administration techniques. Of critical importance in the more recent study is our use of sixteen medical students as research participants. Our goal is to determine 1) whether the more readily available undergraduate students provide data similar to that of actual medical students, and 2) what, if any, unanticipated problems emerge in the administration of our measures to the medical student participants. If medical students perform in a way that is reliably different from our undergraduates, we will use data mining techniques to determine whether there may be important subject selection variables that could be used to recruit groups of undergraduates who behave more like medical students (of great value to future studies evaluating new training technologies in which medical students may not be readily available.).

### *DaVinci Training Study*

In addition to the MIS study, we have also collected performance data on eleven medical students using the DaVinci surgical robot to perform a simple transfer task in both 2D and 3D modes. As with the MIS study, we used our battery of cognitive ergonomics tools to assess mental workload (both subjective metrics and our time estimation secondary-task technique). The goal of this particular study was, once again, to assess the performance of our metrics under progressively more realistic conditions. However, we were also interested in the still-controversial issue of whether 3D imagery provides a significant advantage over more conventional 2D laparoscopic views for

surgical performance. One potentially telling observation that emerged from participant debriefing was that a majority of participants stated that they did not realize that dimensionality was being manipulated during their experimental session. However, their failure to notice the difference between 2D and 3D experimental trials does not guarantee that there will be no performance differences.

We have completed one study in which we compared subjective state measures (workload and stress reports) and time estimation data across three training tasks. Our particular interest was to determine whether using nonmedical student volunteers is a valid means of performing usability tests of, for example, display design variations, or whether medical students must be used. This is a problem of some concern among researchers attempting to evaluate new surgical technologies—some researchers tend to dismiss data collected from nonmedical personnel, and others assume that it is acceptable in some situations. We have argued that the selection of subjects is an appropriate concern, but that for studies of the basic perceptual-motor skills that typify the goals of early training in laparoscopy, students of roughly the same demographics (age/gender) as most introductory medical students will provide roughly equivalent results. For research requiring navigation around simulated organ structures, or actual simulated surgeries, then research participants must have appropriate cognitive models of anatomy and mental scripts appropriate for surgical procedures. In the latter case, it would not be acceptable to substitute research participants of convenience for actual end-users. Thus, generic research participants are not appropriate for all research on the usability of surgical technology, but they are acceptable for studies focusing on facilitation of basic perceptual-motor skills. Our data comparing the two groups of students supported our contention, with one exception. The overall pattern of workload data among tasks was remarkably similar across our two groups of subjects. However, medical students reported higher average workloads. This is not surprising given 1) a likely higher level of motivation and engagement, and 2) a potentially greater tendency to be self-critical among medical students. This means that if absolute levels of workload are important, then more representative (less generic) participants must be recruited. However, when it is the relative performance of participants across conditions that is of inherent interest, as is the case in most usability evaluations, generic subjects seem to provide satisfactory data.

An additional study conducted during the year involved an aspect of workload measurement validation that is rarely seen in the human factors literature, but one that we feel is critically important. If it is assumed that the purpose of workload (and other subjective state measures) is to provide information that can only be obtained by a user carefully evaluating his or her own experience after exposure to a particular apparatus, display, or task, then we would not expect participants who have never performed the task to provide a pattern of results indistinguishable from that obtained from actual post-performance ratings. To address this question, we asked 360 undergraduate student participants to make NASA-TLX ratings of 3 laparoscopic training tasks (cannulation, ring transfer, and cobra rope) based only on static pictures and verbal descriptions similar to those provided as orientation of our full-participation subjects. We then compared the ratings provided by both groups. The pattern of means was not ordered in an identical pattern across both groups, but due to the large variation among subjects in the “control” (no performance) condition, we did not find statistically reliable differences between groups. However, when we compared the mean workload ratings of the two groups, collapsed across specific tasks, we found that the participants who made ratings without having performed the tasks imagined the tasks to have higher workloads than they actually did (based on post-performance reports). This provides some divergent validation of the NASA-TLX as reflecting participants’ perceptions of actual performance rather than simply reflecting performance expectancies. However, the distinctions were not dramatic. We intend to continue exploring this issue as it has not received the attention that it should in usability evaluations, or in the validation of usability metrics in general. We will argue in a paper in preparation that an “expectancy-only” control condition should be employed more frequently when establishing usability evaluation metrics in order to determine whether post-performance measures are, in fact, reflecting the experience of research subjects rather than simple expectancies.

Finally, we have completed collection of data evaluating the utility of binocular disparity cues for research participants using the Da Vinci surgical robot. We ran 10 medical students in July and another 8 in October. After losing four subjects because of a malfunction with one of the robot’s arms in the first session, we were able to obtain enough data to have adequate statistical power to evaluate the hypothesis that the disparity cues reduced workload and stress among participants, even though the participants were unaware of our toggling between 2D and 3D conditions.

## **2. Clinical Assessment at the UMMC MASTRI Center**

Cognitive ergonomic measurements have been performed in research studies to quantify the mental workload associated with the surgeons performing laparoscopic task in different surgical environment including surgical display systems and one-hand and two-hand-surgical techniques.

### *Ergonomic study of surgical display comparison*

Due to the nature of laparoscopic surgery which is minimally invasive, laparoscopic surgeons obtain patients' target anatomy images shown on separate display system. The purpose of this study is to subjectively and objectively compare different surgical display systems with different technologies and image qualities. 10 Laparoscopic surgeons with varying level of surgical experience are recruited for this study and each performs two tasks at height adjustable trainer station. The first task of Fundamentals of Laparoscopic Surgery (FLS), pegboard transfer task, whose goal is to move 6 disks between pegs at the left and right sides, is performed by using two graspers. With a needle hold at a needle driver, surgeons are asked to move two triangles with small eyelets between two circular landings. The image from a laparoscopic camera through s-video cable is fed to three display systems: 1) casually aligned, self calibrating multi-projector display system, 2) size matching single projector, and 3) LCD system. Subjects repeat those two tasks twice with three different display systems in random order. Performance time and number of errors are measured to quantify the quality of their task performance. Through PROMIS surgical simulator, the movements of laparoscopic instruments are measured to quantitatively describe velocity and smoothness.

Mental workload is being assessed by using a subjective and secondary task measure. The subjective measure uses NASA-Task Load Index (NASA-TLX) system. NASA-TLX requires participants to rate their experienced level of workload regarding mental, physical, and time demand, as well as the effort, performance, and frustration during a performance. The subjective assessment on mental workload is conducted immediately after the participants complete a trial using different experimental conditions. Additionally, participants are required to estimate a time length of 21 seconds during laparoscopic training tasks. The accuracy for time estimation is compared between three surgical display systems. The less accurate in time estimation, a higher level of mental workload the surgeon will be recorded. Data collection is in progress.

### *Ergonomic assessment of the difference between one-handed vs. two-handed techniques with virtual reality (VR) simulated laparoscopic cholecystectomy*

It is thought that laparoscopic surgery is a two-handed technique, in which the surgeon who is performing the procedure will be working with the two hands in a kinetic manner. In order to test this hypothesis we are conducting a study in which surgical residents are being monitored while performing a laparoscopic cholecystectomy (LC) on virtual reality (VR) simulators. Each resident is performing the two-handed and one-handed technique separately. The physical workload associated with these two different surgical techniques are analyzed using video monitoring of the laparoscopic view, Electromyography (EMG) analysis of the working muscle groups, with report analysis provided by the simulator at the end of each procedure. Mental workloads are recorded and analyzed by using NASA-TLX and time estimation in the same way described in surgical display comparison study.

## **3. Tools and Display Technology**

**FaceLab:** The team developed custom software to integrate with the FaceLab system via the provided Software Development Kit (SDK). The purpose of this development is to customize the ability to acquire, store and analyze data from the system in the context of the experimental framework we are developing. In particular, we are interested in eye-controlled responses by subjects responding to the NASA TLX questionnaire, which can be seamlessly accomplished with custom code that integrates with the FaceLab device. The current voice-controlled questionnaire can improve on the by-hand collection of questionnaire data, but eye control provides an exceptional level of integration and non-intrusiveness. This follows the overall goal of the STITCH development team, which is working with Dr. Carswell and her team of usability experts in order to automate much of the data collection currently required by the cognitive task studies. During the previous year task complexity measures and subject time estimations were recorded manually, by a research assistant observing the performance of the task. This year the development team spent time researching and testing software/hardware solutions that will allow this data to be captured by software, directly from the subject's verbal utterances. Now we feel that the FaceLab can move this to

an extremely seamless level, even better than speech recognition, using eye control. Overall the goal is to allow for more subjects to be processed with less data entry and quicker analysis.

The FaceLab eye tracking software has also been integrated with our smart display technology. Gaze position can be directly overlaid on a video image in real-time, allowing others to see where a subject's attention is focused. This can be an important training and instructional tool for laparoscopy, as well as allowing our researchers to analyze the differences in techniques between expert surgeons and novices.

**PnP Architecture:** We explored the use of a distributed windowing system (multi-headed X) as a support structure for running a generic and completely flexible windowing environment over the OpenGL-based VIBE display system. Our analysis shows that distributed X will provide the performance and functionality we need and we will move to that environment when several critical OpenGL driver support issues are resolved. We expect this to occur in the near future. In the interim, we have constructed a VNC-based solution that provides a level of performance adequate for functional proof-of-concept demonstrations using our Smart Image client/server system as it is currently implemented. With this system we can run a complete desktop at HD resolution over the system at 10 to 15 frames per second in performance. With optimizations we expect that performance to improve as we anticipate the remaining driver issues to be resolved on the X-server side.

In preparation for the annual Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) conference, we developed a mobile version of our display system. This consisted of building a new display server, at about half the cost and with better performance than the server we constructed last year, and a wheeled cart for mounting our projectors. Further refinements of this system will make an excellent demonstration platform for our technology.

Recent software enhancements include the ability to display full multi-windowed desktop content on the large display, rather than the single application window we have been limited to in the past. This is a key goal in the STITCH display architecture, as it gives us the ability to display all of the information we gather from the operating room environment in a single location, using a scalable, high-resolution, tiled multi-projector array. In our current implementation, latencies in this tiled desktop display are still quite high (approximately 200 ms), but further planned optimizations could bring this to less than 120 ms.

**Creation of a hard real-time development environment for PnP tools development:** Hardware and software have been acquired, configured and passed initial acceptance tests for the creation of a multi-processor hard real-time development system for PnP tools. The system consists of networked InnovationStation IS8200 computing nodes running LinuxWorks LynxOS-178 DO-178B certified hard real-time operating systems and two COTS Dell PC's serving as development workstation and PXE boot domain manager.

We worked closely with LinuxWorks technical support to identify and resolve a number of issues unique to our configuration.

**Identification and assessment of existing medical device interconnection standards:** We have carried out an in-depth analysis of existing medical device interconnection standards and identified a number of standards upon which to base our development work.

- IEEE 11073 is an umbrella standard for a family of medical device interconnection standards. These include domain-specific information models (11073-10101, 11073-10201 and 11073-20101) and domain-specific low-level interconnection models (11073-30200 and 11073-30300).
- CEN EN13735 is an open architecture standard for interconnecting medical devices that has been adapted to the upper and lower layers of IEEE 11073. It defines protocols for establishing, managing and closing connections between medical devices in a master/slave configuration.
- The CIMIT/MGH MD-PnP ICE proposal outlines a more general model of device interconnection in the OR based on sharing of data and control. As such, it focuses on higher-level needs, such as peer-to-peer interconnection, and fault-tolerance at the system level.

The principal complication that arises from attempting to integrate these three standards is their divergent approaches and goals. IEEE 11073 and CEN EN 13735 emphasize data formats, ontology and remote access methods for data without a clear vision for distributed control, safety, reliability or fault tolerance. The CIMIT/MGH MD-PnP ICE proposal emphasizes true peer-to-peer interoperability of networked medical devices without the detailed treatment of data representation or ontology.

At present our plan is a hybrid approach, adopting the IEEE 11073 and EN13735 protocols in an internet-protocol environment. We will use this environment as a test-bed for analysis of ways to extend and adapt these protocols to serve the MD-PnP ICE environment.

**Study Tools:** Our further development activities have involved iterations in the design of the “VisRecognizer,” the voice recognition component of the system that collects spoken responses to the above three measures and codes those responses into a dataset in a spreadsheet format for both immediate analysis and later exploratory analyses. We further integrated this system into our reliability studies using UK’s PSY 100 students, and a user-testing is currently underway at the Vis Center. Results will be compiled into an accuracy matrix for developmental purposes.

- *System Development.* The Visrecognizer is now in its third version. Version 1.0 consisted of computer-synthesized feedback, male voiced questions, and a wide display screen for response feedback. Version 2.0 streamlined the visual display information, utilized both male and female recorded voices for the questions, and implemented human voice feedback. Other implementation changes resulted from a high rate of false positives, false negatives, and confusions (i.e., translating a correctly detected response inaccurately). These changes included: (a) new microphones and soundcards, (2) new microphone position on the user, and (3) filtration adjustment applications for unwanted sound. Version 3.0 (see illustrations) added extra features including: 1) the detection and coding of time estimations in Excel spreadsheet format, 2) microphone configuration options for either a USB microphone or an alternate sound source, 3) implementation of “Pause” and “hold-on” commands, and 4) overall improvements on recognition accuracy.

**Auditory Question Analyzer**

Menu

Question Set:  Intro:

Directory:  /home/dan/Desktop

Filename: NASA\_M\_040708122919.csv

---

Participant #:  Gender: ☒ Male ☐ Female

Block:  Trial:  Task:

Location:

1.) 83  
2.) 64  
3.) 39

Please answer between 0 and 100.

Microphone Status: **Recording** Accuracy: **100%**

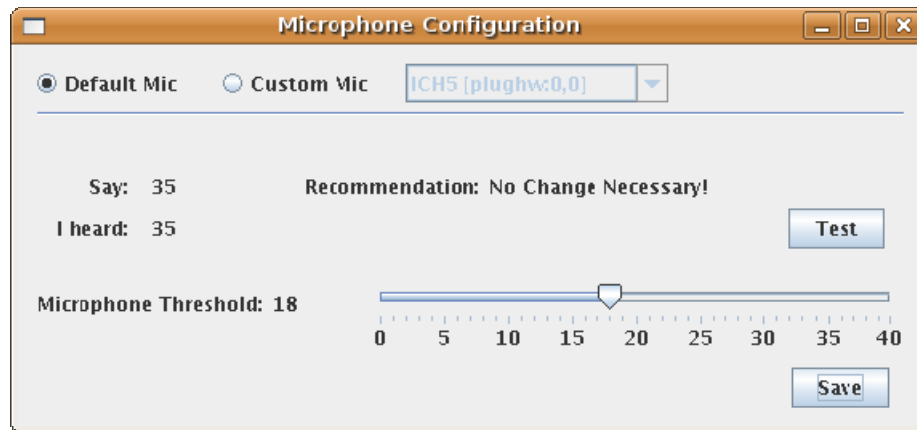
*Screen shot of the VisRecognizer 3.0 showing NASA-TLX responses on the display panel.*

**Time Estimation**

Output File:  /home/dan/timeEstimation.csv

Last Time: 0 secs

*Screen shot of the time estimation display panel.*



*Screen shot of the system's microphone configuration display panel. Adjustment of the microphone threshold allows filtration of background noises.*

- *System integration.* Twenty PSY100 students from UK's Department of Psychology participated in our reliability studies utilizing the VisReconizer Versions 1.0 and 2.0. Five pilot studies were conducted at the Center for Visualization and Virtual Environments and the remainder took place at an additional usability lab on the UK main campus. From students' feedback and experimenters' observations, we modified our experiment procedures to include fewer trials, verbal explanations of procedures to participating students along with a PowerPoint training module.
- *Time estimation automation.* We also modified and began evaluations of a new procedure for administering our time-estimation workload measure. The task now begins with a 21-second tone sample prior to participants' time estimation productions and participants responses are being collected and coded automatically. We will be collecting data from this new system this summer to compare to our experimenter-collected (non-automated) measures from prior experiments.
- *System testing.* We are currently conducting user-testing recruiting staff and students from the Vis Center. The goal is to test the VisRecognizer's accuracy rates in recognizing of voices of different genders and accents. Data will be compiled into an accuracy matrix.

Our latest software tool for studying cognitive human factors has been deployed in studies with University of Kentucky students. VocalTLX is a speech recognition program for automating the collection of subjective workload data during experimental trials. The NASA-TLX, MRQ, and SSSQ questionnaires are all integrated in VocalTLX, with automatic auditory prompts and automated recognition of the subject's responses. Trials are currently underway to determine the accuracy of the system, and to compare its results to past studies.

**SAGES:** In April, the STITCH project was featured in the Learning Center of the 2008 Meeting of the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) in Philadelphia. We presented the overall vision of the integrated STITCH environment, using our high-resolution multi-projector display technology, integrated instrumentation, and automated workload measurement with speech recognition and eye tracking. Our booth was one of the larger attractions in the Learning Center, at about 20 feet by 20 feet, and had good foot traffic throughout the exhibition sessions at SAGES. Dr. Brent Seales presented his work in surgical visualization during one of the scientific sessions. Dr. Cindy Lio and research assistants Matt Field and Scott Ross staffed the booth, and gathered feedback on our work from the attendees of the conference.

The booth was arranged with four major stations. One featured information on our research, in the form of several posters and handouts, while the media for research documentary by Steve Bailey highlighted the key areas of research and the personnel involved. In the center of the booth, an array of six projectors created a seamless tiled, blended display of nearly 4 megapixels. This display demonstrated our vision of an integrated operating environment, by providing a common area to show multiple, simultaneous information feeds. Live video from an endoscope was combined on-screen with real-time eye-tracking information from the subject, and displayed alongside telemetry information from sensors mounted on surgical instruments.

In parallel, a mock display based on the work of Dr. Qiong Han showed a pre-operative reconstruction of internal anatomy based on CT scans, illustrating our vision of an enhanced surgical experience by providing more relevant and more tightly integrated information to the surgeon. The goal of Dr. Han's dual display software is to simultaneously show a real-time endoscopic video feed and a corresponding reconstruction based on a combination of pre-operative scans and known anatomical models. The camera view can be registered to the model to produce a three-dimensional "map" of the surgical field to guide the procedure. The software is still at an early stage of development, but we were able to demonstrate a simulation of the final product, using computer-generated imagery rather than live video. Dr. Lio used this version to gather user feedback on the design and to establish additional requirements for a better surgical experience.

The potential for the use of eye tracking to study cognitive workload generated much enthusiasm from the attendees. Several surgeons showed interest in future collaboration, offering their time and facilities to expand our subject pool for human factors study. At another station, Dr. Lio demonstrated our workload measurement tools, particularly our speech recognition software, and interviewed surgeons on their reactions to using it. We plan to use the feedback gathered from these interviews to guide the next iteration of project tools. The final station was staffed by our collaborators from the MASTRI Center at the University of Maryland Medical Center. They presented the Maryland Visual Comfort Index and surveyed attendees about endoscopic image quality.

We would especially like to thank Dr. Adrian Park of UMMC for the opportunity to demonstrate our work at SAGES this year.

#### **4. Supported Personnel**

The following is a list of personnel receiving pay from the research effort.

Catherine M Carswell	Scott Ross
James D Hoskins	Eric L. Coker
William B Seales	Russell C. Grant
Charles Pike	Ryan F. Baumann
Dorothy Porter Leontseva	Daniel L Staley
Stephen Strup	Qiong Han
Stephen P Bailey	Robert T. Tipton
Danny S Crasto	Cara E. Worick
Matthew Douglas Field	Michelle Sublette

### **Key Research Accomplishments (Summary)**

- Reliability assessments for equivalent-forms reliability of printed and spoken versions of the NASA-TLX.
- Reliability assessment of the voice recognition software program.
- Assessment of the performance of medical students and evaluation of equivalence with psychology students for STITCH study purposes.
- Collection of performance data for eleven medical students using the DaVinci surgical robot to perform a simple transfer task in both 2D and 3D modes.
- Ergonomic study of surgical display comparison in the UMMC MASTRI center.
- Ergonomic assessment of the difference between one-handed vs. two-handed techniques with virtual reality (VR) simulated laparoscopic cholecystectomy.
- Creation of the VisRecognizer tool for hands-free administration of the NASA-TLX questionnaire.
- Creation of a hard real-time development environment for PnP tools development.
- Identification and assessment of existing medical device interconnection standards.
- Software to integrate with the FaceLab system with the NASA-TLX questionnaire.

# Reportable Outcomes

## ABSTRACT ACCEPTED FOR AN ORAL PRESENTATION

- Park A, Lee G, Meenaghan N, Lee TH, Seagull FJ (2008) Patients benefit while surgeons suffer: an impending epidemic, Annual meeting of American College of Surgeons (ACS)
- Lee G, Lee TH, Dexter DJ, Godinez C, Meenaghan N, Park AE (2008) Joint kinetic data augments traditional biomechanical approach to assess the ergonomics of laparoscopic camera assistants, Annual conference of Medicine Meets Virtual Reality (MMVR), Long Beach, CA

## FULL PAPER ACCEPTED FOR AN ORAL PRESENTATION

- Carswell, C.M., Lio, C., Grant, R., Seales, B., and Clarke, D. (2008). Equivalent-forms reliability of printed and spoken versions of the NASA-TLX. Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society. Santa Monica, CA: The Human Factors and Ergonomics Society.

## ABSTRACT SUBMITTED FOR REVIEW

- Lee G, Meenaghan N, Lee TH, Dexter DJ, Godinez C, Seagull FJ, Park A (2008) A pain in the neck! The relationship of video monitors to surgeons's stress, Annual conference of SAGES 2009, Phoenix, AZ
- Lee G, Lee TH, Dexter DJ, Park AE (2008) Traditional biomechanical approach augmented by joint kinetic data in ergonomic risk assessment of laparoscopic assistants, Annual conference of SAGES 2009, Phoenix, AZ

## MANUSCRIPT IN PREPARATION

Park A, Lee G, Seagull FJ, Carlos Godinez, Meenaghan N, Lee TH (2008) Patients benefit while surgeons suffer: an impending epidemic, Annals of Surgery

# Conclusion

During project year 2 the STITCH project has made significant progress in all of its objective areas. (1) Assessment of task cognitive ergonomics by psychology researchers has yielded significant results, some published, many yet to be published, in demonstrating the validity of our methods. These methods have also been applied to realistic studies using medical students and actual clinical equipment (the DaVinci robot). (2) Significant progress was made in the deployment of study tools in the MASTRI center at the University of Maryland Medical Center. These tools were applied to the study of medical students and residents in the MASTRI center, yielding significant, published results. And (3), enabling standards and technologies have been identified for the creation of the plug-and-play tools environment, a hard real-time development environment created, and work has been initiated on the design and development of prototype tools for this environment.

Remaining project objectives will be completed under a no-cost extension granted through 15 May 2009. This includes significant activity focused on publication of the results of completed experiments, as well as software development activity in support of the VisRecogniser tool, the plug-and-play environment, and prototype plug-and-play tools.